NOTICE: When government or other drawings, specifications or other data are used for any purpose other than in connection with a definitely related government procurement operation, the U. S. Government thereby incurs no responsibility, nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use or sell any patented invention that may in any way be related thereto.

62-4-5

D1-82-0190

BOEING SCIENTIFIC RESEARCH LABORATORIES

Numerical Procedures for Tchebycheff Approximation

E. W. Cheney

Mathematics Research

June 1962

D1-82-0190

NUMERICAL PROCEDURES FOR TCHEBYCHEFF APPROXIMATION

by

Professor E. W. Cheney
University of California, Los Angeles
(Visiting Staff Member)
Summer 1962

Mathematical Note No. 264

Mathematics Research Laboratory

BOEING SCIENTIFIC RESEARCH LABORATORY

June, 1962

These notes were prepared by Professor E. W. Cheney for a series of lectures on "Numerical Methods for Approximation" at the Boeing Scientific Research Laboratories in June, 1962.

OUTLINE OF LECTURES

I. Examples of problems to be considered. A one-dimensional example for introductory purposes. The reduction of many approximation problems to an overdetermined system of linear equations. Possibility of non-uniqueness in minimax solutions.

II. Theoretical background. Convexity. Linear inequalities. Characterizing the solution of minimax problems. Solution of systems involving $n + 1$ equations in n unknowns.

III. The exchange theorem. An ascent algorithm for the minimax problem. Formulas for computation. Using the algorithm to solve linear inequalities.

IV. Rational approximation problem. Existence of best approximations. Pitfalls. Changing rational functions into continued fraction form for fast computing. A linear inequality method for rational approximations. A weighted minimax algorithm.

V. The differential correction algorithm. Padé approximations. Examples.

LECTURE I

We begin by explaining what is meant by a Tchebycheff approximation. A simple example, taken from the book "Approximations for Digital Computers" by Cecil J. Hastings (Princeton University Press, 1955) is as follows:

$$\text{Arctan } x \approx c_1 x + c_2 x^3 + c_3 x^5 + c_4 x^7$$

$$\begin{aligned} c_1 &= .9992150 & & \\ c_2 &= -.3211819 & \epsilon &= .00008 \\ c_3 &= .1462766 & \text{for } & 0 \le x \le 1 \\ c_4 &= -.0389929 & & \end{aligned}$$

The number ϵ is the maximum discrepancy on the interval $[0,1]$ between Arctan x and the polynomial approximation. This alone would not justify the appellation "Tchebycheff approximation". The crucial fact is that the number ϵ can not be improved (decreased) by any adjustments in the coefficients given above. That is, we have reduced the number

$$\epsilon = \max_{0 \le x \le 1} | \text{Arctan } x - (c_1 x + c_2 x^3 + c_3 x^5 + c_4 x^7) |$$

to an absolute minimum by chosing the coefficients $c_1, \ldots, c_4$ as shown. In this brief course we shall develop methods for the numerical determination of the coefficients in such a Tchebycheff approximation. Our techniques are not at all restricted to <u>polynomial</u> approximation, however. Their scope is illustrated by the following typical problems to which they apply.

(1) Find a polynomial $P(x)$ of lowest degree such that on the interval $[0,\frac{\pi}{4}]$, $|P(x) - \sin x| \leq 10^{-8} \sin x$.

(2) Find a polynomial in two variables of the form $P(x,y) = \sum_{i+j\leq 4} c_{ij} x^i y^i$ such that $\max_{\substack{|x|\leq 1\\ |y|\leq 1}} |f(x,y) - P(x,y)|$ is a minimum, f being a prescribed function.

(3) Find a rational function $R(x)$ of lowest total degree such that on the interval $[0,1]$, $|R(x) - \text{Arctan}\, x| \leq 10^{-16}$.

(4) Given a function $f(x)$ which is known only at certain points $x_1,\ldots,x_m$, find a polynomial $P(x)$ of degree 5 for which the expression

$$\max_{1\leq i\leq m} |f(x_i) - P(x_i)|$$

is an absolute minimum.

(5) Given an overdetermined system of linear equations

$$\sum_{j=1}^{n} a_{ij} x_j = d_i \qquad (i = 1,\ldots, m)$$

find an approximate solution $x = (x_1,\ldots, x_n)$ for which the expression

$$\max_{1\leq i\leq m} \left| \sum_{j=1} a_{ij} x_j - d_i \right|$$

is an absolute minimum.

(6) Continuous functions f , g_1 , $g_2, \ldots, g_n$ being prescribed, find the best approximation of f by a linear combination of $g_1, \ldots, g_n$ on an interval $[a,b]$. That is, determine coefficients $c_1, \ldots, c_n$ in such a way that the expression

$$\max_{a \leq x \leq b} \left| f(x) - \sum_{i=1}^{n} c_i g_i(x) \right|$$

shall be an absolute minimum.

The algorithms which we will develop subsequently are capable of handling all these problems. The most fundamental of these problems is number (5) , the approximate solution of overdetermined linear equations. For practical purposes many approximation problems may be put into this form. For example, suppose we wish to calculate the coefficients in the approximation cited earlier:

$$\text{Arctan } x \approx c_1 x + c_2 x^3 + c_3 x^5 + c_4 x^7 \ .$$

On the interval $[0,1]$ let us take a large number of points $x_1, \ldots, x_m$. We then wish to determine the coefficients $c_1, \ldots, c_4$ so as to minimize the expression

$$\max_{1 \leq i \leq m} \left| \text{Arctan } x_i - (c_1 x_i + c_2 x_i^3 + c_3 x_i^5 + c_4 x_i^7) \right| \ .$$

If we put $a_{ij} = x_i^{2j-1}$ and $d_i = \text{Arctan } x_i$, then we seek to minimize

$$\max_{1 \leq i \leq m} \left| \sum_{j=1}^{4} a_{ij} c_j - d_i \right| \ ,$$

which is an instance of problem (5). One may object that we have replaced one problem by another and that the solution of the second need not be close to a solution of the original. However, it is possible to prove under suitable hypotheses that the approximation taken on a finite set of points approaches the approximation for the interval, as the finite set "fills out" the interval. The interested reader may consult my Boeing Document No. D1-82-0185, "The Relationship between the Tchebycheff Approximations on an Interval and on a Discrete Subset of that Interval", for a discussion of this problem.

For reasons set forth above, we are going to consider first the problem of minimizing an expression

$$\Delta(x) = \max_{1 \leq i \leq m} \left| \sum_{j=1}^{n} a_{ij} x_j - d_i \right|$$

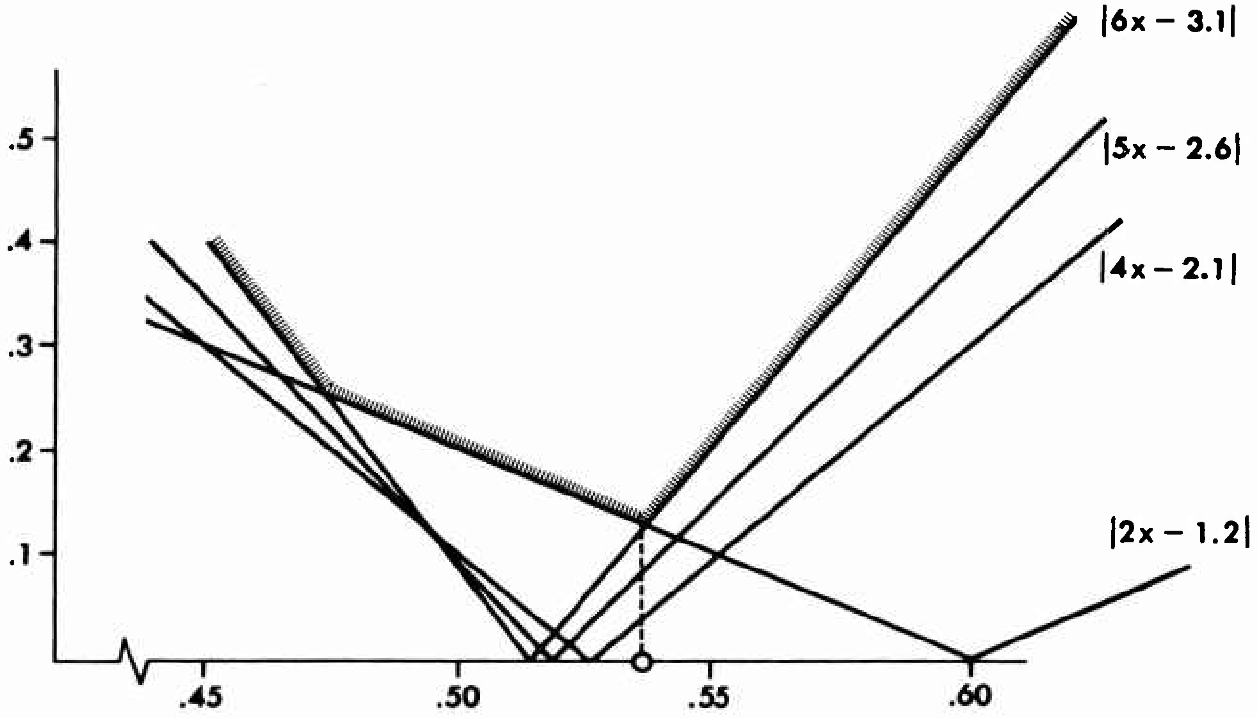
where the data a_{ij} and d_i are prescribed (real) numbers. In order to see what to expect, let us examine a simple example, in which $n = 1$ and $m = 4$:

$$\begin{aligned} 2x &= 1.2 \\ 4x &= 2.1 \\ 5x &= 2.6 \\ 6x &= 3.1 \quad . \end{aligned}$$

It is clear that any approximate solution to this system of equations should lie between .5 and .6 . We seek to minimize the function

$$\Delta(x) = \max \{ |2x - 1.2| , |4x - 2.1| , |5x - 2.6| , |6x - 3.1| \} .$$

We shall do this graphically in order to gain some insight into the nature of the function Δ . We first plot the functions $|2x - 1.2|$ etc., and then select the topmost curve, in accordance with the operation of taking the maximum.



From the graph we see that the solution can be calculated exactly by locating the intersection of two lines. We want one of the two solutions of the equation

$$|2x - 1.2| = |6x - 3.1| ,$$

the other being a spurious solution. The correct one is obtained from

$$1.2 - 2x = 6x - 3.1$$

and is therefore $x = .5375$.

Several features of this example that will presist when n and m are much greater deserve to be pointed out. First of all, the solution is obtainable by solving a simple linear equation, but a relatively great expenditure of effort went into the discovery of this equation. Secondly, the solution is a point where two of the residual functions, defined by

$$\begin{aligned} r_1(x) &= 2x - 1.2 \\ r_2(x) &= 4x - 2.1 \\ r_3(x) &= 5x - 2.6 \\ r_4(x) &= 6x - 3.1 \end{aligned}$$

are equal in magnitude. (In the general case the solution will be a point where $n + 1$ of the residuals are equal in magnitude.) Finally, the minimum point of our function is the same as the minimum point of the simpler function

$$\Delta_1(x) = \max \{ |2x - 1.2| , |6x - 3.1| \} .$$

The numerical implications of this observation are quite important. A large value of m will not make the computations unstable or ill-conditioned; it will simply involve a higher number of iterations to locate the appropriate system of $n + 1$ equations which determines the solution. Those who are familiar with least-squares computations will realize that this is an important advantage. The least-squares solution of a matrix equation

$$A x = d$$

is obtained as the exact solution of

$$A^T A x = A^T d \ .$$

Simply forming $A^T A$ when m is large may involve great round-off errors. Generally speaking, a Tchebycheff solution of a system of equations can be calculated with higher precision than a least-squares solution.

It should be remarked that a Tchebycheff solution of an over-determined system of linear equations need not be unique. For example, the Tchebycheff solutions of the system

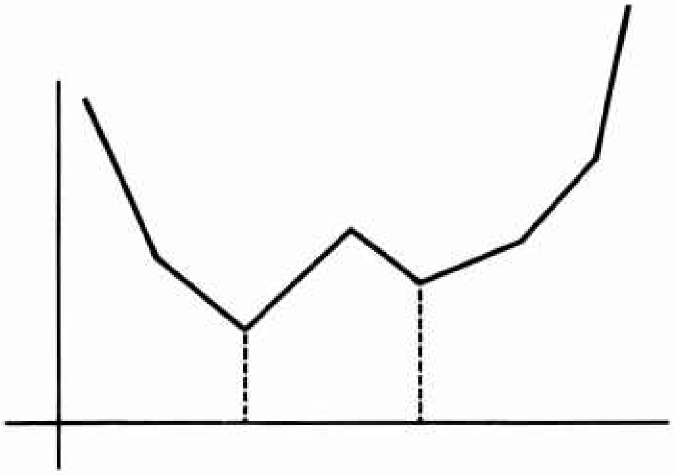
$$\left\{ \begin{aligned} x &= 1 \\ 0\,x &= 1/2 \\ 2\,x &= 1 \end{aligned} \right.$$

fill out an interval $[\frac{1}{2}, \frac{3}{4}]$. In the general case this can occur only if some $n \times n$ submatrices from A are singular.

Proceeding now to the general problem of minimizing the function

$$\Delta(x) = \max_{1 \leq i \leq m} \left| \sum_{j=1}^{n} a_{ij} x_j - d_i \right| ,$$

we shall prove that Δ cannot have any purely <u>local</u> minima, that is, it cannot have a graph such as the following.



For, if possible, let x and y be two local minimum points of Δ. For $\theta \varepsilon [0,1]$ we find that

$$\Delta[\theta x + (1 - \theta) y] = \max_i \left| \Sigma a_{ij} [\theta x_j - (1 - \theta) y_j] - d_i \right|$$

$$\leq \theta \max_i \left| \Sigma a_{ij} x_j - d_i \right| + (1 - \theta) \max_i \left| \Sigma a_{ij} y_j - d_i \right|$$

$$= \theta \Delta (x) + (1 - \theta) \Delta (y) .$$

(Thus the function Δ is convex.) If $\Delta(x) < \Delta(y)$, then some points near y will have lower values of Δ than y. This can be seen by taking values of θ near 0. On the other hand, if $\Delta(x) = \Delta(y)$, then there can be no higher points of the graph between them since $\Delta [\theta x + (1 - \theta) y] \leq \Delta (x)$. A similar proof would apply to the more complicated situation

$$\Delta (c) = \max_{a \leq x \leq b} \left| f(x) - \sum_{i=1}^{n} c_i g_i (x) \right| .$$

The implication of this fact for the problem considered earlier,

$$\text{Arctan } x \approx c_1 x + c_2 x^3 + c_3 x^5 + c_4 x^7$$

is that, if no "infinitesimal" alteration of the coefficients can decrease the maximum discrepancy ϵ , then no "massive" alteration of the coefficients will do so either.

It should be pointed out that the problem of locating the minimum point of the function

$$\max_i \left| \sum_j a_{ij} x_j - d_i \right|$$

can be solved by "linear programming". To do so we introduce another variable ϵ and ask that it be a minimum under the conditions

$$\Sigma a_{ij} x_j - d_i \leq \epsilon$$

$$- \Sigma a_{ij} x_j + d_i \leq \epsilon \quad .$$

This is a standard problem of minimizing a linear function with linear inequality constraints.

LECTURE II

A basic concept for what follows is that of a convex set. A set K is convex if

$$\left.\begin{array}{l} x \in K \\ y \in K \\ 0 \leq \lambda \leq 1 \end{array}\right\} \Longrightarrow \lambda x + (1 - \lambda) y \in K$$

Thus with any two points x , y in K , the line segment joining x and y lies also in K . Given any set U , another set $\mathcal{H}(U)$ is determined by specifying that it contain all linear combinations

$$\Sigma \lambda_i u_i$$

in which $u_i \in U$, $\lambda_i \geq 0$, and $\Sigma \lambda_i = 1$. The number of summands in the linear combination is arbitrary (but always finite). The set $\mathcal{H}(U)$ is easily shown to be convex and is called the convex hull of U . The convex hull of three points, for example, is the triangle having those points as vertices. An important theorem states that in an n-dimensional linear space, the sum $\Sigma \lambda_i u_i$ can be restricted to just $n + 1$ terms.

Theorem of Caratheodory In an n-dimensional space every point of $\mathcal{H}(U)$ can be written $x = \sum_{i=0}^{n} \lambda_i u_i$ where $u_i \in U$, $\lambda_i \geq 0$, and $\Sigma \lambda_i = 1$.

Proof If $x \in \mathcal{H}(U)$, then $x = \sum_{i=0}^{k} \lambda_i u_i$ with $\lambda_i \geq 0$, $u_i \in U$, and

$\Sigma\lambda_i = 1$. Let us assume that k is <u>minimal</u>, yet $k > n$, and try to obtain a contradiction. Since $\sum_0^k \lambda_i(x - u_i) = 0$ the points $y_i = x - u_i$ are linearly dependent. Since we are assuming $k > n$, the points $y_1, \ldots, y_k$ are also linearly dependent, say $\sum_1^k \alpha_i y_i = 0$. Since k was minimal , $\lambda_i > 0$. Put $\alpha_0 = 0$. Clearly

$$\sum_0^k (\lambda_i + t\alpha_i) y_i = 0$$

for all t . When $t = 0$, the coefficients are <u>positive</u>. As we increase $|t|$ the coefficients remain positive for a while but eventually one will vanish. They don't <u>all</u> vanish because $\lambda_0 + t\alpha_0 = \lambda_0 > 0$. But if we are careful to take the first t , at least one coefficient will be 0 , while those that are not 0 are positive. Going back to x and u , we contradict the minimality of k, in as much as $\Sigma(\lambda_i + t\alpha_i)(x - u_i) = 0$ whence $x = \Sigma(\lambda_i + t\alpha_i)u_i / \Sigma(\lambda_i + t\alpha_i)$. The division at the end makes the coefficient add up to 1 .

<u>Theorem on Linear Inequalities</u> A system of linear inequalities

$$\sum_{j=1}^n a_{ij} x_j > 0 \qquad i = 1, \ldots, m$$

is consistent if and only if $0 \notin \mathcal{K}\{A_1, \ldots, A_m\}$. Here A_i denotes the n-tuple $(a_{i1}, \ldots, a_{in})$.

<u>Proof</u> If the convex set $K = \mathcal{K}(A_1, \ldots A_m\}$ does not contain the origin, let x be the point of K closest to 0 . Then for any i , and for any $\lambda \in [0,1]$, $\lambda A_i + (1 - \lambda)x \in K$. Consequently

$0 \leq \| \lambda A_i + (1-\lambda) x \|^2 - \| x \|^2 = \| \lambda (A_i - x) + x \|^2 - \|x\|^2 =$ $\lambda^2 \| A_i - x \|^2 + 2\lambda \langle A_i - x, x \rangle$. But this inequality would be violated for small $\lambda > 0$ unless $\langle A_i - x, x \rangle \geq 0$. Thus $\langle A_i, x \rangle \geq \langle x, x \rangle > 0$. This shows that x is a solution of the system of inequalities. For the converse, assume $0 \in K$. Then $0 = \sum_{i=1}^{m} \lambda_i A_i$ with $\lambda_i \geq 0$ and $\Sigma \lambda_i = 1$. Thus $0 = \langle 0, x \rangle = \Sigma \lambda_i \langle A_i, x \rangle$. This cannot be true if all $\langle A_i, x \rangle > 0$. The notation $\langle u, v \rangle$ denotes $\sum_{j=1}^{n} u_j v_j$.

Now we return to the problem of minimizing the function

$$\Delta(x) = \max_{1 \leq i \leq m} \left| \sum_{j=1}^{n} a_{ij} x_j - d_i \right| .$$

How will we know when a point $x = (x_1, \ldots, x_n)$ is a solution? It must be impossible to decrease Δ by moving slightly in any direction from x . Among the residuals

$$r_i(x) = \sum_j a_{ij} x_j - d_i \qquad i = 1 \ldots m$$

let us single out those which are maximum in absolute value. By renumbering the original equations we could assume

$$|r_1(x)| = |r_2(x)| = \ldots = |r_k(x)| > |r_{k+i}(x)| \qquad (i = 1, 2, \ldots) .$$

Thus k is the number of residuals which are equal in absolute value to $\Delta(x)$, and we have assumed that these are the first k . Now if we are to decrease $\Delta(x)$ by changing x , we will have to decrease all the numbers

$$|r_1(x)|, \ |r_2(x)| , \ \ldots , \ |r_k(x)| .$$

The remaining residuals which are less can be ignored if only a slight change is contemplated in x . Suppose that we move from x in the direction z . How are the residuals affected? A computation shows:

$$r_i(x + \lambda z) = \Sigma a_{ij}(x_j + \lambda z_j) - d_i = \Sigma a_{ij} x_j - d_i + \lambda \Sigma a_{ij} z_j =$$
$$= r_i(x) + \lambda \langle A_i, z \rangle .$$

Thus r_i increases when $\langle A_i, z \rangle$ is positive; it decreases if $\langle A_i, z \rangle$ is negative; and it remains constant if $\langle A_i, z \rangle = 0$. In order to decrease $|r_i(x)|$ by moving in direction z , then, $\langle A_i, z \rangle$ should be of opposite sign to $r_i(x)$. Let us define $\sigma_i = \operatorname{sgn} r_i(x)$, so that $\sigma_i = +1, 0, -1$ according as $r_i(x) > 0, = 0, < 0$. The direction z that we are seeking must then have the property

$$\sigma_1 \langle A_1, z \rangle < 0 , \quad \sigma_2 \langle A_2, z \rangle < 0 , \ldots , \sigma_k \langle A_k, z \rangle < 0 .$$

In other words, z must be a solution of the system of linear inequalities

$$\langle \sigma_i A_i , z \rangle < 0 \qquad (i = 1, \ldots, k) .$$

At any point, x , we can define such a system of inequalities by singling out the residuals which are maximum in absolute value and letting σ_i denote the sign of these residuals. If this system of linear inequalities is <u>consistent</u>, then x is not a solution, for a slight displacement of x in an appropriate direction z will decrease $\Delta(x)$. If the system of linear inequalities is <u>inconsistent</u>, then there is no direction in which all the maximum residuals decrease and hence x is a solution.

Let us assume as a non-degeneracy hypothesis that every set of n

vectors selected from the set $\{A_1, \dots, A_m\}$ is independent. An equivalent hypothesis is that every $n \times n$ submatrix from $A = (a_{ij})$ is non-singular. Under this hypothesis we can prove that at the solution point x there must be at least $n + 1$ maximum residuals in absolute value. Suppose that x is a point at which there are n or fewer maximum residuals, say $r_1, \dots, r_k$, where $k \leq n$. Then we can solve the system of equations

$$\langle A_i, z \rangle = - \sigma_i \qquad . \quad i = 1, \dots, k$$

because of our non-degeneracy assumption, and the resulting direction z will be one in which Δ decreases, because $\langle \sigma_i A_i , z \rangle < 0$. (Recall that in the simple example with $n = 1$ there were two equal maximum residuals in absolute value at the minimum point.) In the above argument the case $\sigma_i = 0$ does not arise unless all $r_i = 0$, when there will be m equal maximum residuals.

In preceding theorems, we have shown that, if the system of linear inequalities

$$\langle A_i, z \rangle < 0 \qquad i = 1, \dots, k$$

is to be inconsistent, then $0 \in \mathcal{H}\{A_1, A_2, \dots, A_k\}$. We have also shown that 0 must lie in the convex hull of no more than $n + 1$ of the points $A_1, \dots, A_k$. We have therefore proved the following theorem:

Theorem A minimax solution of a system of m equations in n unknowns, where $m > n$ is a minimax solution of a certain subsystem comprising $n + 1$ equations.

The next thing to do then is to see how to compute the minimax solution of a system of $n + 1$ equations in n unknowns. For convenience in notation let the system be written

$$\langle A_i , x \rangle = d_i \qquad (i = 0, \ldots, n) .$$

Suppose that by some means we are able to obtain a point $x = (x_1, \ldots, x_n)$, a number ϵ, and signs σ_i $(-1, 1)$ in such a way that the two following conditions are satisfied.

$$(1) \qquad \langle A_i , x \rangle - d_i = \sigma_i \epsilon \qquad (i = 0, \ldots, n)$$

$$(2) \qquad 0 \in \mathcal{X}\{\sigma_0 A_0, \sigma_1 A_1, \ldots, \sigma_n A_n\} .$$

Then we would be finished, for by (1), all $n + 1$ residuals are equal to $|\epsilon|$ in magnitude, and by (2), no reduction in the number

$$\Delta(x) = \max_{0 \leq i \leq n} | \langle A_i , x \rangle - d_i |$$

is possible. Probably the easiest way to obtain x, ϵ, and σ_i is as follows.

<u>First</u>, obtain any non-trivial solution $(\lambda_0, \ldots, \lambda_n)$ of the homogeneous equations

$$\sum_{i=0}^{n} \lambda_i A_i = 0 .$$

<u>Second</u>, define $\sigma_i = \operatorname{sgn} \lambda_i$ $(i = 0, \ldots, n)$. If any λ_i vanishes, then the non-degeneracy assumption has been violated. Since $\Sigma \sigma_i \lambda_i \sigma_i A_i = 0$

and $\sigma_i \lambda_i > 0$, we have already secured condition (2) above.

Third, put $\epsilon = -\Sigma \lambda_i d_i / \Sigma \lambda_i \sigma_i$.

Fourth, solve for x in the system

$$\langle A_i, x \rangle = d_i + \sigma_i \epsilon \qquad (i = 1, \ldots, n) \quad .$$

This system is solvable because it is n equations in n unknowns, with a non-singular matrix. We almost have fulfilled condition (1). We must merely check to verify that

$$\langle A_0, x \rangle = d_0 + \sigma_0 \epsilon \quad .$$

To do this compute as follows:

$$\sum_0^n \lambda_i r_i(x) = \Sigma . \lambda_i [\langle A_i, x \rangle - d_i] = \langle \Sigma \lambda_i A_i, x \rangle - \Sigma \lambda_i d_i =$$

$$= - \Sigma \lambda_i d_i = \epsilon \Sigma \lambda_i \sigma_i = \epsilon \lambda_0 \sigma_0 + \sum_1^n \lambda_i \epsilon \sigma_i =$$

$$= \epsilon \lambda_0 \sigma_0 + \sum_1^n \lambda_i r_i(x) \quad .$$

Cancelling, we get

$$\lambda_0 r_0(x) = \epsilon \lambda_0 \sigma_0$$

which was to be proved.

Remark 1 If the signs σ_i were known, then we could treat condition (1) as a system of $n + 1$ linear equations in the unknowns $x_1, \ldots, x_n, \epsilon$ and be done immediately. This is usually possible in polynomial approximation problems.

Remark 2 If we assume only that *some* set of n vectors from $\{A_0, \ldots, A_n\}$ is independent (rather than *all* sets), then the only change necessary in the above algorithm is in solving for x , where we would have to *select* a set of n rows with a non-singular matrix in order for Gaussian elimination to work smoothly.

LECTURE III

We now return to the more general problem of m equations in n unknowns. The minimax solution of the whole system is also the minimax solution of an $(n + 1)$ - subsystem, and we must systematize the search for this subsystem. Let us start by calculating the minimax solution of any subsystem

$$\langle A_{i_j}, x \rangle = d_{i_j} \qquad (j = 0,\ldots, n) \quad ,$$

where the indices $i_0, i_1,\ldots, i_n$ are any $n + 1$ taken from the set $\{1, 2,\ldots, m\}$. Having the point x , we now compute all residuals

$$r_i(x) = \langle A_i, x \rangle - d_i \quad .$$

Of course, $n + 1$ of these should be equal in magnitude, but unless we are very lucky they will not be the <u>maximum</u> ones. If $|r_{i_0}(x)|,\ldots,|r_{i_n}(x)|$ were the maximum residuals, then x would be the minimax solution of the entire system. So we select an index α such that

$$|r_\alpha(x)| = \Delta(x) = \max_{1 \leq i \leq m} |r_i(x)|$$

and we are now going to replace one of the indices $i_0,\ldots, i_n$ by α and repeat the entire process. Strangely enough, the index i_j which is to be replaced is uniquely determined by the condition that

$$0 \in \mathcal{H}\{\sigma_\alpha A_\alpha, \sigma_{i_0} A_{i_0},\ldots, \sigma_{i_{j-1}} A_{i_{j-1}}, \sigma_{i_{j+1}} A_{i_{j+1}},\ldots, \sigma_{i_n} A_{i_n}\} \ .$$

Here as usual we have put $\sigma_i = \operatorname{sgn} r_i(x)$. The formal statement of this fact is known as the exchange theorem.

<u>Exchange Theorem</u> In E_n let $\{A_0, \ldots, A_{n+1}\}$ be a set of vectors of which every set of n is independent. Suppose that 0 is in the convex hull of $\{A_0, \ldots, A_n\}$. Then there is a unique index β such that 0 is in the convex hull of $\{A_0, \ldots, A_{\beta-1}, A_{\beta+1}, \ldots, A_{n+1}\}$.

<u>Proof</u> By hypothesis there exist constants $\theta_i \geq 0$ such that $0 = \sum_{i=0}^{n} \theta_i A_i$ and $\Sigma \theta_i = 1$. Since every set of n vectors is independent, $\theta_i > 0$. We may express A_{n+1} as a linear combination $A_{n+1} = \sum_{i=0}^{n} \mu_i A_i$, and all possible expressions of 0 as a linear combination of $A_0, \ldots, A_{n+1}$ are encompassed by the equation

$$s \left[A_{n+1} - \sum_{i=0}^{n} (\mu_i - t\theta_i) A_i \right] = 0$$

where t and s are real variables. If $s > 0$, then for large t the coefficients $-\mu_i + t\theta_i$ are all positive, and for an appropriate value of t , all of these coefficients are positive save one which vanishes. (Specifically, we take t equal to the largest ratio μ_i / θ_i .) No more than 1 coefficient vanishes at a time for otherwise there is a contradiction of the hypothesis concerning independence. The index of the vanishing coefficient is therefore uniquely determined if we require that the remaining coefficients be positive.

This theorem is due to E. Stiefel, "Numerical Methods of Tchebycheff Approximation", pp. 217-231 in the book "On Numerical Approximation", R. Langer, editor, Madison, 1959. The reader should consult also E. Stiefel, "Note on Jordan Elimination, Linear Programming, and Tchebycheff Approximation", Numerische Mathematik 2 (1960) 1-17.

In broad outline our algorithm is this:

1. Given any set of $n + 1$ indices, $I = \{i_0, \ldots, i_n\}$, calculate the Tchebycheff solution of the $n + 1$ equations $\sum_{j=1}^{n} a_{i_k j} x_j = d_{i_k}$ $(k = 0, \ldots, n)$. Let the point obtained be $x = (x_1, \ldots, x_n)$.

2. Let α be an index for which $|r_\alpha(x)| = \Delta(x)$.

3. Perform an "exchange" of α with an appropriate index from I. Having this new set of indices, I, return to step 1.

In detail the algorithm is as follows.

1. Select any set of indices $I = \{i_0, \ldots, i_n\}$. By Gaussian elimination, calculate any non-trivial solution $(\lambda_0, \ldots, \lambda_n)$ of the homogeneous equations $\sum_{k=0}^{n} a_{i_k j} \lambda_k = 0$ $(j = 1, \ldots, n)$. Define $\sigma_k = \operatorname{sgn} \lambda_k$ $(k = 0, \ldots, n)$. If any $\lambda_k = 0$, note this fact. Define

$$\epsilon = - \sum_{k=0}^{n} \lambda_k d_{i_k} / \sum_{k=0}^{n} |\lambda_k| .$$

By Gaussian elimination, calculate the solution $x = (x_1, \ldots, x_n)$ of the equations $\sum_{j=1}^{n} a_{i_k j} x_j = d_{i_k} + \sigma_k \epsilon$ $(k = 1, \ldots, n)$. Test to see whether $\sum_{j=1}^{n} a_{i_0 j} x_j - d_{i_0} - \sigma_0 \epsilon = 0$. Record this number.

2. Calculate the numbers $r_i(x) = \sum_{j=1}^{n} a_{ij} x_j - d_i \quad (i = 1, \ldots, m)$.
Select α so that $|r_\alpha(x)|$ is a maximum. Put $\sigma_\alpha = \operatorname{sgn} r_\alpha(x)$.
If $|r_\alpha(x)| \leq |\epsilon|$, stop, for x is the Tchebycheff solution.

3. By Gaussian elimination, calculate the solution $\mu = (\mu_1, \ldots, \mu_n)$ of the equations $\sum_{k=1}^{n} \mu_k \sigma_k a_{i_k j} = \sigma_\alpha a_{\alpha_j} \quad (j = 1, \ldots, n)$. Define $\mu_0 = 0$. Let β be the index of the largest ratio $\mu_k / \sigma_k \lambda_k$. Replace i_β by α , and return to step 1 with the new set of indices I .

There are some methods of streamlining the above computations to save a little computer time. But I believe the above algorithm to be superior in maintaining accuracy throughout the calculation.

It is an interesting fact that any algorithm for solving over-determined systems of linear equations can be used without modification to solve another type of problem, viz. solving systems of linear inequalities. A system of linear inequalities looks typically like this:

$$\sum_{j=1}^{n} a_{ij} x_j \leq d_i \qquad (i = 1, \ldots, m) \quad . \tag{1}$$

We have here a system of m inequalities in n unknowns, and we seek a solution x , if one exists. If one exists, the system is said to be consistent. We have the following theorem:

Theorem If the system of linear inequalities above is consistent, then for sufficiently large constants Q, every Tchebycheff solution of

$$\sum_{j=1}^{n} a_{ij} x_j = d_i - Q \qquad (i = 1,\ldots, m) \qquad (2)$$

is a solution of the inequalities.

<u>Proof</u>. Let y be any solution of the inequalities. Let $Q \geq \max_i (d_i - \Sigma a_{ij} y_j)$. Let x be a Tchebycheff solution of (2). Then

$$\max_i (\Sigma a_{ij} x_j - d_i + Q) \leq \max_i |\Sigma a_{ij} x_j - d_i + Q| \leq$$

$$\leq \max_i |\Sigma a_{ij} y_j - d_i + Q| = \max_i (\Sigma a_{ij} y_j - d_i + Q) \leq Q ,$$

whence $\Sigma a_{ij} x_j \leq d_i$. A similar result for least-squares solutions is <u>not</u> true.

LECTURE IV

By the "rational approximation problem" we mean the following. Let a continuous function f defined on an interval $[a,b]$ be given. Let n and m be two prescribed integers. Determine then the optimum coefficients $p_0, \ldots, p_n$, $q_0, \ldots, q_m$ in the approximation

$$f(x) \approx \frac{p_0 + p_1 x + p_2 x^2 + \ldots + p_n x^n}{q_0 + q_1 x + q_2 x^2 + \ldots + q_m x^m} .$$

As usual we would like a Tchebycheff solution, and thus seek to minimize the expression

$$\Delta = \max_{a \leq x \leq b} \left| f(x) - \frac{P(x)}{Q(x)} \right| ,$$

where P and Q are respectively the numerator and denominator of the rational function. Of course, there is a discrete analogue of this problem in which we select a finite set of points from the interval $[a,b]$.

The existence of the optimum coefficients is guaranteed by a theorem in Achieser, Theory of Approximation, Ungar, 1956, p. 53 . (The proof has a flaw but is essentially correct.) It is necessary to emphasize that the existence is provable for best rational approximations on an interval but not for a discrete subset! For example, suppose f is a continuous function defined on $[0,2]$ such that $f(0) = 1$, $f(1) = 0$, and $f(2) = 0$. Let us attempt to approximate f by a rational function of the form $a/(bx + c)$ at just the points 0, 1 and 2 . Now the deviation of the function $\epsilon/(x + \epsilon)$ from f on the three given points is $\epsilon/(1 + \epsilon)$, and this can be made arbitrarily small by making ϵ small. Thus

$$\inf_{a,b,c} \max_{x=0,1,2} \left| f(x) - \frac{a}{bx + c} \right| = 0 \ .$$

Yet no choice of a, b, and c will achieve for us this minimum. This example is due to P. C. Curtis, Jr. Another phenomenon that may occur in the discrete problem is that after obtaining a good rational approximation for the discrete point set, one may discover that the denominator vanishes at intermediate points, so rendering the approximation useless. For example, let us try to approximate $f(x) = x$ on $[-1,1]$ by $a/(bx + c)$. On the subset

$$\underline{\overline{X}}_n = \{ x : \frac{1}{n} \leq |x| \leq 1 \}$$

the best approximation of the stated form is $\frac{1}{nx}$, but this has a pole at $x = 0$. This remains true as $n \to \infty$. This example is also due to Curtis. Thus certain pitfalls await the unwary in this field.

One may ask whether, in view of these drawbacks, rational approximations are worth our study. After all, polynomials are sufficient for the approximation of any continuous function - so says the Weierstrass Theorem. Nevertheless, a polynomial approximation to a given continuous function may fail to have the required accuracy unless its degree becomes very large. In these cases a rational approximation will sometimes provide a spectacular improvement. From the standpoint of economizing the computing time, rational functions are also recommended. The reason for this is that a rational function can always be converted into an equivalent continued fraction for fast computing. To illustrate, consider the rational function

$$R(x) = \frac{x^4 + 2x^3 - 2x^2 - 2x + 4}{x^3 + 2x^2 + 2x + 6} \quad .$$

It would appear that eight "long operations" (multiplications and divisions) would be necessary to compute $R(x)$. However, if we perform the long division indicated, we get

$$R(x) = x + \frac{-4x^2 - 8x + 4}{x^3 + 2x^2 + 2x + 6} \quad .$$

Now write this as

$$R(x) = x - \frac{4}{\dfrac{x^3 + 2x^2 + 2x + 6}{x^2 + 2x - 1}} \quad .$$

We again perform the indicated long division to obtain

$$R(x) = x - \frac{4}{x + \dfrac{3x + 6}{x^2 + 2x - 1}}$$

or

$$R(x) = x - \frac{4}{x - \dfrac{3}{\dfrac{x^2 + 2x - 1}{x + 2}}} \quad .$$

Continuing in this way we obtain finally

$$R(x) = x - \frac{4}{x\ -}\ \frac{3}{x\ -}\ \frac{1}{x + 2}$$

which obviously entails only 3 long operations. Nevertheless, the rational function that we started with has roughly the same curve-fitting ability as a polynomial of degree seven or eight.

One of the simplest algorithms for obtaining rational approximations depends on having at hand a program for solving linear inequalities. (In this connection see remarks made in lecture 3.) Suppose we wish an approximation of the form

$$f(x) \approx \frac{\sum_{j=1}^{n} c_j x^{j-1}}{1 + \sum_{n+1}^{N} c_j \; x^{j-n}} \quad .$$

Let us require that the approximation shall deviate from f no more than an amount ϵ at a large number of points $x_1, x_2, \ldots, x_m$. Let us require also that the denominator remain $\geq \delta > 0$ at these m points. Our requirements are then

$$\left| f(x_i) - \frac{\sum_{j=1}^{n} c_j \; x_i^{j-1}}{1 + \sum_{n+1}^{N} c_j \; x_i^{j-n}} \right| \leq \epsilon$$

$$1 + \sum_{n+1}^{N} c_j \; x_i^{j-n} \geq \delta \quad .$$

This becomes a system of linear inequalities if we replace an inequality $|y| \leq \epsilon$ by two inequalities $y \leq \epsilon$, $-y \leq \epsilon$, and then clear of fractions. The result is

$$\left\{\begin{aligned} [f(x_i) - \epsilon] \sum_{n+1}^{N} c_j x_i^{j-n} - \sum_{1}^{n} c_j x_i^{j-1} &\leq \epsilon - f(x_i) \\ [-f(x_i) - \epsilon] \sum_{n+1}^{N} c_j x_i^{j-n} + \sum_{1}^{n} c_j x_i^{j-1} &\leq \epsilon + f(x_i) \\ - \sum_{n+1}^{N} c_j x_i^{j-n} &\leq 1 - \delta \quad . \end{aligned}\right.$$

If we select too small an ϵ this system will have no solution $(c_1, \ldots, c_n, \ldots, c_N)$. But by trial and error, we can discover an appropriate value of ϵ and a satisfactory rational approximation.

Another simple and very effective method which can be used has been termed the Weighted Minimax Algorithm. It requires that one already possess a program for solving overdetermined systems of linear equations in the minimax or Tchebycheff sense. This algorithm suffers from the disadvantage that a proof of its effectiveness is still lacking. Nevertheless, the ease with which it may be programmed and the success which it has had recommend it highly. To explain it, suppose again that we seek an approximation

$$f(x) \approx \frac{P(x)}{Q(x)}$$

where P and Q are polynomials of prescribed degrees whose coefficients we wish to determine. We seek, of course, to render the following expression an absolute minimum:

$$\max_{a \le x \le b} \left| f(x) - \frac{P(x)}{Q(x)} \right| .$$

This can be written

$$\max_{a \le x \le b} \left| \frac{1}{Q(x)} \right| . \left| f(x)\, Q(x) - P(x) \right| .$$

Now in this algorithm we consider $\left| \frac{1}{Q(x)} \right|$ to be a weight function, $w(x)$. In the first step we take $w(x) \equiv 1$, and attempt to minimize

$$\max_{a \le x \le b} \left| f(x)\, Q(x) - P(x) \right| .$$

To avoid trivial solutions, we may fix a term in $Q(x)$, say by requiring $Q(0) = 1$. Then we compute a new weight function $w(x)$ and repeat the process. Formally, we have the following iterative algorithm.

1. Using $w_i \equiv 1$ minimize the expression $\max\limits_{1 \le i \le m} w_i \left| f(x_i)\, Q(x_i) - P(x_i) \right|$ by use of the program for overdetermined linear equations. Call the solution Q_0 and P_0.

2. Set $w_i = 1/|Q_0(x_i)|$ and minimize the expression $\max\limits_{i} w_i \left| f(x_i)\, Q(x_i) - P(x_i) \right|$. Call the solution P_1, Q_1.

3. Set $w_i = 1/|Q_1(x_i)|$ and continue in this way.

Experience has shown that using double precision arithmetic and about 10 steps in the above algorithm we can usually obtain best Tchebycheff rational approximations with up to 15 coefficients.

LECTURE V

The next algorithm to be discussed is perhaps not as easily programmed as the Weighted Minimax Algorithm but rests on firmer ground mathematically. Again we are attempting to optimize the coefficients in two polynomials P and Q so that

$$f(x) \approx \frac{P(x)}{Q(x)} \qquad \text{on} \quad [a,b] \quad .$$

We find it convenient to define $R(c,x) = P(c,x)/Q(c,x)$,

$$P(c,x) = \sum_{i=1}^{n} c_i \, x^{i-1} \quad , \qquad Q(c,x) = \sum_{i=n+1}^{N} c_i \, x^{i-n-1}$$

and

$$\Delta(c) = \max_{a \le x \le b} \left| f(x) - \frac{P(c,x)}{Q(c,x)} \right| \quad .$$

The letter c stands for the N-tuple $(c_1, \ldots, c_n, \ldots, c_N)$. Since numerator and denominator can be multiplied by the same number without changing the rational function, no loss of generality occurs in restricting c to the cube

$$\overline{K} = \{ c \in E_n : \max_i |c_i| \le 1 \} \quad .$$

Algorithm The N-tuple c^0 may be arbitrary except that $Q(c^0,x) > 0$ in $[a,b]$. This gets the process started. Now at any stage, with N-tuple c^ν on hand, define an auxilliary function

$$\delta_\nu(c) = \max_{a \le x \le b} \{ |f(x)\,Q(c,x) - P(c,x)| - \Delta(c^\nu)\,Q(c,x) \} \quad .$$

Notice that the function remains substantially the same throughout the computation; it changes only in the single coefficient $\Delta(c^{\nu})$. Now select $c^{\nu+1}$ to minimize this auxilliary function on the cube K . (This is a problem of "convex programming" - minimizing a convex function on a convex set.) If $\delta_{\nu}(c^{\nu+1}) < 0$, repeat the whole process. If $\delta_{\nu}(c^{\nu+1}) \geq 0$, stop, for c^{ν} was a solution.

<u>Theorem</u> As $\nu \to \infty$, $\Delta(c^{\nu}) \downarrow \inf\limits_{c\ \ E_N} \Delta(c)$.

<u>Proof</u> A. If $Q(c^{\nu+1}, x_0) \leq 0$ ever occurs for $x_0 \in [a,b]$, let ν be the first index for which it occurs. Thus $Q(c^{\nu+1}, x_0) \leq 0$ but $Q(c^{\nu},x) > 0$ for all x . We show c^{ν} is a solution. If not, then there is a c' such that $\Delta(c') < \Delta(c^{\nu})$. We may assume $c' \in K$, and that $Q(c',x) > 0$ in $[a,b]$, for if $Q(c',x)$ vanishes anywhere in $[a,b]$ we can divide a common factor out of numerator and denominator. Thus $\delta_{\nu}(c^{\nu+1}) \leq \delta_{\nu}(c') =$
$= \max\limits_{x} \{[\ |f(x) - R(c',x)| - \Delta(c^{\nu})\ ]\ Q(c',x)\} < 0$. But $\delta_{\nu}(c^{\nu+1}) \geq$
$\geq |f(x_0)\, Q(c^{\nu+1},x_0) - P(c^{\nu+1},x_0)| - \Delta(c^{\nu})\ Q(c^{\nu+1},x_0) \geq 0$, a contradiction.

B. We assume $Q(c^{\nu},x) > 0$ always. We shall show that $\delta_{\nu}(c^{\nu+1}) \leq 0$, equality occurring only if c^{ν} is a solution. Indeed, $\delta_{\nu}(c^{\nu+1}) \leq \delta_{\nu}(c^{\nu}) =$
$= \max\limits_{i} \{[|f(x) - R(c^{\nu},x)| - \Delta(c^{\nu})]\ Q(c^{\nu},x)\} = 0$. If c^{ν} is not a solution then as in part A above, we can show that $\delta_{\nu}(c^{\nu+1}) < 0$.

C. $\Delta(c^{0}) > \Delta(c') > \ldots$. To prove this, write $0 > \delta_{\nu}(c^{\nu+1}) =$
$= \max\limits_{x} \{[|f(x) - R(c^{\nu+1}, x)| - \Delta(c^{\nu})]\ Q(c^{\nu+1}, x)\} \geq$
$\geq \beta\,[\max\limits_{x} |f(x) - R(c^{\nu+1}, x)| - \Delta(c^{\nu})] = \beta\,[\Delta(c^{\nu+1}) - \Delta(c^{\nu})]$ where
$\beta = \max\limits_{c \in K}\ \max\limits_{a \leq x \leq b}\ Q(c,x)$.

D. Put $\Delta^* = \inf_{c \in K} \Delta(c)$. Put $L = \lim_{\nu \to \infty} \Delta(c^\nu)$. The number L is well-defined because it is the limit of a decreasing sequence of non-negative numbers. If $L \neq \Delta^*$, then there is a $c' \in K$ such that $\Delta(c') < L$. We may assume that $Q(c',x) > 0$ on $[a,b]$. Then $|f(x) - R(c',x)| \leq \Delta(c') < L \leq \Delta(c^\nu)$, so that $\delta_\nu(c^{\nu+1}) \leq \delta_\nu(c') = $ $= \max_x \{[|f(x) - R(c',x)| - \Delta(c^\nu)]\, Q(c',x)\} \leq \alpha[\Delta(c') - \Delta(c^\nu)]$ where $\alpha = \min_{a \leq x \leq b} Q(c',x) > 0$. Thus

$$\Delta(c^{\nu+1}) \leq \frac{1}{\beta}\, \delta_\nu(c^{\nu+1}) + \Delta(c^\nu)$$

$$\leq \frac{\alpha}{\beta}\, [\Delta(c') - \Delta(c^\nu)] + \Delta(c^\nu) \quad .$$

Now in this last inequality, let $\nu \to \infty$. The result is

$$L \leq \frac{\alpha}{\beta}\, [\Delta(c') - L] + L$$

whence $0 \leq \Delta(c') - L$, a contradiction. This concludes our proof. This algorithm occurs in E. W. Cheney and H. L. Loeb, "On Rational Chebyshev Approximation", Numerische Mathematik, Summer, 1962. The convex programming problem is discussed, among other places, in E. W. Cheney and A. A. Goldstein, "Newton's Method for Convex Programming and Tchebycheff Approximation", Numerische Mathematik 1 (1959) 253-268.

The last type of rational approximation which we wish to discuss is the so-called Padé approximation. Let f be an analytic function, given by a Taylor series

$$f(x) = \sum_{k=0}^{\infty} a_k x^k \quad .$$

A rational function

$$\frac{P_m(x)}{Q_n(x)} = \frac{\sum_0^m p_k x^k}{\sum_0^n q_k x^k}$$

is a Padé approximation to f if $Q_n f - P_m$ is approximately zero, in the sense that it has a power series in which the first $n + m + 1$ coefficients vanish:

$$Q_n(x) f(x) - P_m(x) = \sum_{k=n+m+1}^{\infty} c_k x^k \quad .$$

For an example, let us determine a Padé approximation for e^x of the form

$$e^x \approx \frac{p_0 + p_1 x}{q_0 + q_1 x + q_2 x^2} \quad .$$

We must compute p_0, p_1, q_0, q_1, q_2 in accordance with

$$(q_0 + q_1 x + q_2 x^2)(1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots) - (p_0 + p_1 x) = c_4 x^4 + c_5 x^5 + \dots$$

Equating coefficients of like powers of x, we have

$$q_0 - p_0 = 0$$

$$q_0 + q_1 - p_1 = 0$$

$$\frac{1}{2} q_0 + q_1 + q_2 = 0$$

$$\frac{1}{3} q_0 + \frac{1}{2} q_1 + q_2 = 0 \quad .$$

A convenient solution is

$$e^x \approx \frac{6 + 4x}{6 - 2x + x^2} .$$

Putting into continued fraction form, we have

$$e^x \approx \frac{4}{x - \frac{7}{2} +} \; \frac{\frac{45}{4}}{x + \frac{3}{2}} .$$

The general procedure is as follows.

$$\left(\sum_{k=0}^{n} q_k x^k\right)\left(\sum_{k=0}^{\infty} a_k x^k\right) - \sum_{k=0}^{m} p_k x^k = \sum_{k=n+m+1}^{\infty} c_k x^k .$$

We multiply the series together and combine:

$$\sum_{k=0}^{\infty} \left[\sum_{j=0}^{k} a_{k-j} q_j - p_k \right] x^k = \sum_{n+m+1}^{\infty} c_k x^k .$$

Again equate coefficients to obtain the linear equations:

$$\sum_{k=0}^{k} a_{k-j} q_j - p_k = 0 \qquad (k = 0, 1, \ldots, n+m)$$

$$p_k = 0 \quad \text{when} \quad k > m$$

$$q_k = 0 \quad \text{when} \quad k > n$$

$$q_0 = 1 .$$

The last equation simply helps us to settle upon a particular solution of the homogeneous equations. Sometimes we are forced to specify some coefficients other than q_0 . A formula for the coefficients c_k is

$$c_k = \sum_{j=0}^{n} a_{k-j} q_j \qquad (k > m + n) .$$

Since

$$f(x) - \frac{P_m(x)}{Q_n(x)} = \frac{\sum_{k>n+m} c_k x^k}{Q_n(x)}$$

the c's help in assessing the error. For example in the case of e^x with $m = 1$ and $n = 2$ we have for $k > 3$,

$$c_k = a_k q_0 + a_{k-1} q_1 + a_{k-2} q_2$$

$$= \frac{6}{k!} - \frac{2}{(k-1)!} + \frac{1}{(k-2)!}$$

$$= \frac{(k-3)(k-2)}{k!} \quad .$$

Thus $c_4 = \frac{1}{12}$, $c_5 = \frac{1}{20}$, $c_6 = \frac{1}{60}$ etc. The error function is therefore

$$\frac{\Sigma c_k x^k}{Q_n(x)} = \frac{\frac{1}{12}x^4 + \frac{1}{20}x^5 + \frac{1}{60}x^6}{x^2 - 2x + 6} \quad .$$

When x is near zero, the denominator is near 6 , and so the error is of the order of $\frac{1}{72}x^4$. Remember that this is obtained from an expression which involves just two divisions. In the general case, taking $n = m$, we would have an error of the order of x^{2n+1} at the expense of n divisions - two times better than the Taylor series.